Graph-based data mining for biological applications

Leander Schietgat
Department of Computer Science,
Katholieke Universiteit Leuven,
Celestijnenlaan 200A, 3001 Leuven, Belgium
E-mail: leander.schietgat@cs.kuleuven.be

In many real-world problems, one deals with input or output data that are structured. This thesis investigates the use of graphs as a representation for structured data and introduces relational learning techniques that can efficiently process them. We apply the techniques to two biological problems. On the one hand, we use decision trees to predict the functions of genes, of which the hierarchical relationships can be structured as a graph. On the other hand, we predict chemical activity of molecules by representing them as graphs. We show that, by exploiting graph properties, efficient learning techniques can be developed. It turns out that in both cases, the relational models are not only learned more efficiently, but their predictive performance significantly improves as well.

Keywords: structured prediction, hierarchical multi-label classification, graph mining, structure-activity learning

1. Introduction

During the last decade, high-throughput techniques such as gene sequencing, microarray experiments and molecular screenings have generated huge amounts of biological and chemical data. The availability of these data has created opportunities for data mining, which can extract useful knowledge from them. Despite many successful applications, most data mining methods require propositional data, while many biological data such as DNA sequences, proteins or molecules are structured.

The ability to handle structured data is one of the challenges that is being tackled by relational data mining. The overall goal of this thesis [2] is to improve the efficiency of relational learning techniques, as well as their applicability to problems from biology and chemistry. We will try to achieve this goal by representing the data as graphs and exploiting their specific properties.

2. Decision trees for hierarchical multi-label classification

In the first part, we study the task of hierarchical multi-label classification (HMC), a variant of classification where an example may belong to multiple classes and where the classes are organized in a hierarchy. A key application of HMC is gene function prediction. It is known that a gene may have multiple functions, while biologists have organized these functions into hierarchies. Instead of following an approach that learns an independent model for each class, we propose an approach that learns a single model predicting all classes at once. The output of the model then consists of a graph representing the functions and their relationships.

Our motivation for using decision trees is that they are efficiently learnable on large datasets and that they lend themselves to interpretation by domain experts. Moreover, the technique to learn them, known as top-down induction of decision trees, can easily be adapted to the HMC setting. First, we show that HMC decision trees are not only learned more efficiently than binary decision trees, but they are also superior in terms of model size and interpretability. Perhaps the most surprising result is that HMC decision trees also obtain a higher predictive performance, which can be explained by the fact that they are less prone to overfitting [5]. Second, we show that an ensemble version of HMC trees leads to a similar boost in predictive performance as found in binary classification tasks. Finally, we have shown that our ensemble HMC method outperforms the state-of-the-art methods for gene function prediction on three model organisms in terms of efficiency, predictive performance (see e.g., Fig. 1) and usability [4].

AI Communications
ISSN 0921-7126, IOS Press.

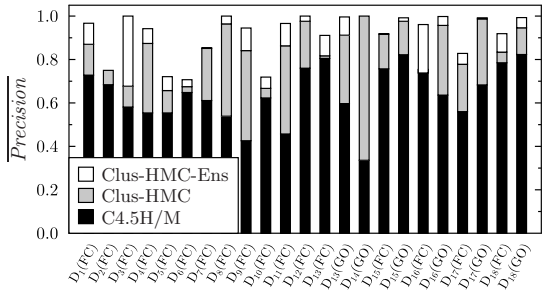

Fig. 1. Comparison of average precision between C4.5H/M (a state-of-the-art HMC decision tree learner discussed in [4]), Clus-HMC and Clus-HMC-Ens, at the recall obtained by C4.5H/M, on 23 datasets from functional genomics.

In conclusion, HMC decision trees are state-of-the-art models for gene function prediction and should definitely be considered in HMC tasks where interpretable models are desired.

3. Maximum common subgraph mining

In the second part, we consider learning and mining tasks in which the input data of the learning algorithms are represented as graphs. The application we focus on is the learning of structure-activity relationships (SAR), where the goal is to predict properties of molecules based on their atom-bond structure.

In order to make the learning algorithms more efficient, we exploit specific properties of molecular graphs. Motivated by the fact that the majority of molecules can be represented as outerplanar graphs and that the block-and-bridge-preserving (BBP) subgraph isomorphism is a suitable matching operator in the SAR context, we propose a polynomial algorithm that computes a maximum common subgraph (MCS) of two outerplanar graphs. The intuition behind the MCS of two molecular graphs is that it may contain shared properties relating to their chemical activity.

On the one hand, we show that it is possible to construct an intuitive metric for molecules based on the MCS. Moreover, the metric is efficiently computable and outperforms state-of-the-art metrics in terms of predictive performance [1]. On the other hand, we use MCSs as a feature representation for graphs [3]. The feature generation method is efficient to compute, does not have parameters that need to be tuned, results in a smaller, non-redundant feature set and obtains a state-of-the-art predictive performance on several SAR tasks. Table 1 shows a comparison between mining all MCS features (A-MCS), random MCS features (R-MCS), all subgraph features (A-GP), chemical fingerprints (FP2) and correlated subgraph features (C-GP). The results indicate that the MCS feature generation methods reach a similar or better predictive performance while using less patterns.

Table 1
Average AUROC scores and ranks on 60 NCI datasets for the state-of-the-art feature generation methods. The Friedman test is used to compute statistical significance [3].

Method	Average AUROC	Average rank
A-MCS	0.796	1.45
A-GP	0.796	1.55
R-MCS	0.784	3.18
FP2	0.779	3.82
C-GP	0.684	5
Critical difference for the average ranks at 1%: 1.36		

In conclusion, we have proposed two graph mining techniques for the classification of molecules and we have shown that the BBP subgraph isomorphism should be preferred for SAR tasks, because of its efficiency and chemical relevance.

Acknowledgements Leander Schietgat was supervised by Hendrik Blockeel and Maurice Bruynooghe and supported by IWT and ERC SG 240186.

References

[1] L. Schietgat, J. Ramon, M. Bruynooghe, and H. Blockeel. An efficiently computable graph-based metric for the classification of small molecules. In *Proc. of the 11th International Conference on Discovery Science* (LNAI 5525), pp. 197–209, 2008.

[2] L. Schietgat. *Graph-based data mining for biological applications.* PhD thesis, Department of Computer Science, K.U.Leuven, 2010. URL: https://lirias.kuleuven.be/handle/ 123456789/267094.

[3] L. Schietgat, F. Costa, J. Ramon, and L. De Raedt. Effective feature construction by maximum common subgraph sampling. *Machine Learning*, 2010. DOI: 10.1007/s10994-010-5193-8.

[4] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Džeroski. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinf.*, 11(2), 2010.

[5] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.